CLAIMS

1. A method for load balancing a plurality of servers, the method comprising:

providing a plurality of control blocks, each control block associated with a number of active connections a server is connected with, the control block configured to control at least one server with the associated number of connections in a server list;

causing each control block to point to a server with a least value ascertained by determining the number of connections on the server relative to the server's capacity to handle connections;

selecting the control block associated with the least number of connections; and

selecting the server pointed to by the control block.

2. The method as in claim 1, wherein ascertaining the least value for the server comprises:

determining a metric of the server by dividing the number of connections on the server by the capacity of the server, wherein the metric is kept as a quotient/remainder pair;

storing the quotient/remainder pair in the control block;

incrementing the remainder by one for every connection added to the server; and decrementing the remainder by one for every connection removed from the server.

3. The method as in claim 1, further comprising:

causing the control block with the server having an added/removed connection to transfer the server to an adjacent control block, wherein the adjacent control block is associated with the number of connections pertaining to the transferring server;

causing the control block to transfer the metric of the server to the adjacent control block; and

updating the pointer to point to the next server on the list of the control block.

4. The method as in claim 3, further comprising:
removing the control block if the control block does not have a server on the server list.

5. The method as in claim 3, further comprising:

causing the adjacent control block to receive the transferring server;

causing the adjacent control block to receive the metric of the transferring server; and

causing the adjacent control block to update and sort the server list.

6. The method as in claim 5, further comprising:

adding a control block if there is no control block associated with the number of connections of the transferring server.

7. A processor executable medium which when executed by a processor performs a method for load balancing a plurality of servers, the method comprising:

providing a plurality of control blocks, each control block associated with a number of active connections a server is connected with, the control block configured to control at least one server with the associated number of connections in a server list;

causing each control block to point to a server with a least value ascertained by determining the number of connections on the server relative to the server's capacity to handle connections;

selecting the control block associated with the least number of connections; and selecting the server pointed to by the control block.

8. The processor executable medium as in claim 7, wherein ascertaining the least value for the server comprises:

determining a metric of the server by dividing the number of connections on the server by the capacity of the server, wherein the metric is kept as a quotient/remainder pair;

storing the quotient/remainder pair in the control block;

incrementing the remainder by one for every connection added to the server; and decrementing the remainder by one for every connection removed from the server.

9. The processor executable medium as in claim 7, further comprising:

causing the control block with the server having an added/removed connection to transfer the server to an adjacent control block, wherein the adjacent control block is associated with the number of connections pertaining to the transferring server;

causing the control block to transfer the metric of the server to the adjacent control block; and

updating the pointer to point to the next server on the list of the control block.

10. The processor executable medium as in claim 9, further comprising:

removing the control block if the control block does not have a server on the server list.

11. The processor executable medium as in claim 9, further comprising:

causing the adjacent control block to receive the transferring server;

causing the adjacent control block to receive the metric of the transferring server; and

causing the adjacent control block to update and sort the server list.

12. The processor executable medium as in claim 11, further comprising:

adding a control block if there is no control block associated with the number of connections of the transferring server.

13. A load balancing apparatus comprising:

a plurality of control blocks, each control block associated with a number of active connections a server is connected with, the control block configured to control at least one server with the associated number of connections;

a pointer in each control block that points to a server with a least value ascertained by determining the number of connections on the server relative to the server's capacity to handle connections; and

a selection circuit that selects the control block associated with the least number of connections and further selects the server pointed to by the control block.

14. The load balancing apparatus as in claim 13, further comprising:

The control block configured to determine a metric of the server, wherein the metric is kept as a quotient/remainder pair;

a memory to store the quotient/remainder pair determined by the control block;

the control block further configured to increment the remainder by one for every connection added to the server; and

the control block further configured to decrement the remainder by one for every connection removed from the server.

15. The load balancing apparatus as in claim 13, further comprising:

the control block configured to transfer the server having an added/removed connection to an adjacent control block, wherein the adjacent control block is associated with the number of connections pertaining to the transferring server;

the control block further configured to transfer the metric of the server to the adjacent control block; and

the control block configured to update the pointer to point to the next server on the list of the control block.

16. The load balancing apparatus as in claim 15 further comprises:

the control block is de-activated if the control block does not have a server on the server list.

17. The load balancing apparatus as in claim 15, further comprises:

the adjacent control block configured to receive the transferring server; and

the adjacent control block further configured to receive the metric of the transferring server, wherein the adjacent control block updates and sorts the server list.

18. The load balancing apparatus as in claim 17, further comprises:

a control block that is activated to receive the transferring server if there is no control block associated with the number of connections of the transferring server and the control block is associated with the number of connections of the transferring server.